National Aeronautics and Space Administration

NASA GeneLab: Open Science for Life in Space

Biological and Physical Sciences

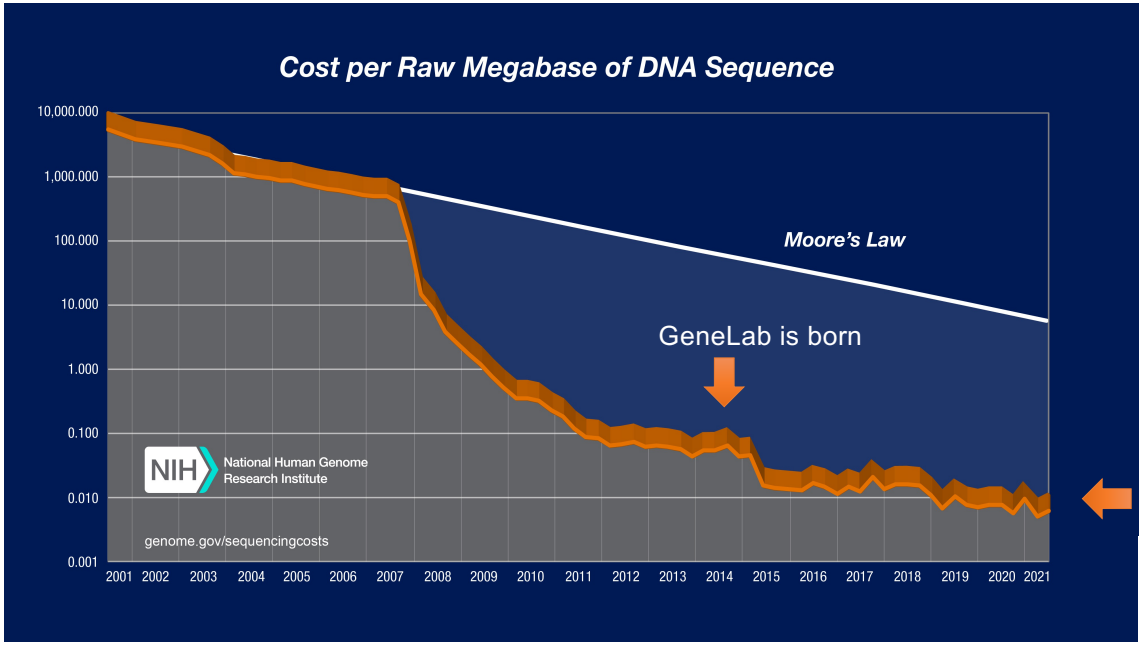
Jonathan Oribello
GeneLab Data Processing Team
Bioinformatics Programmer
NASA Ames Research Center
Contractor: Blue Marble Space

Presentation Overview

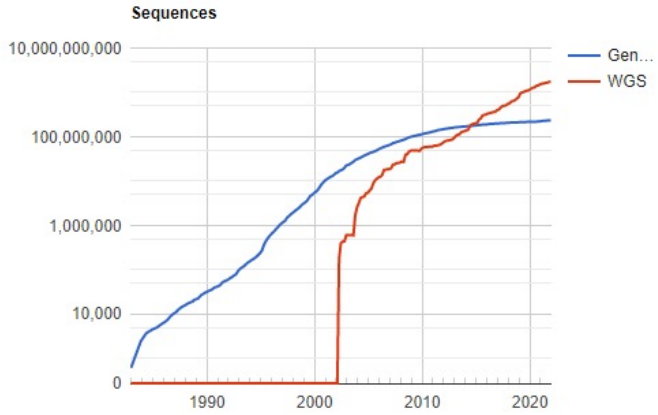
Topics Covered

- Omics approach for space biology research
- Overview of GeneLab
- Data Processing Goals and Current Work

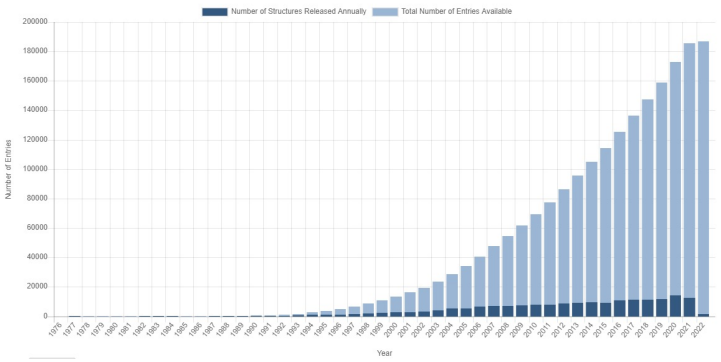
The Sequencing Paradigm Shift: Mountains of Data



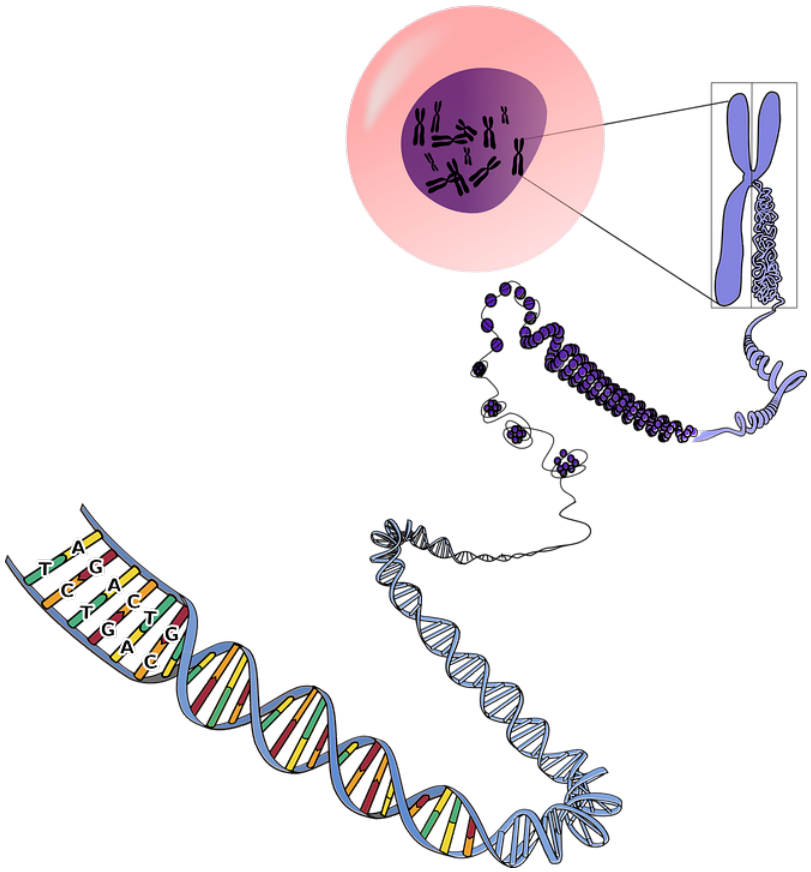
GenBank Sequences

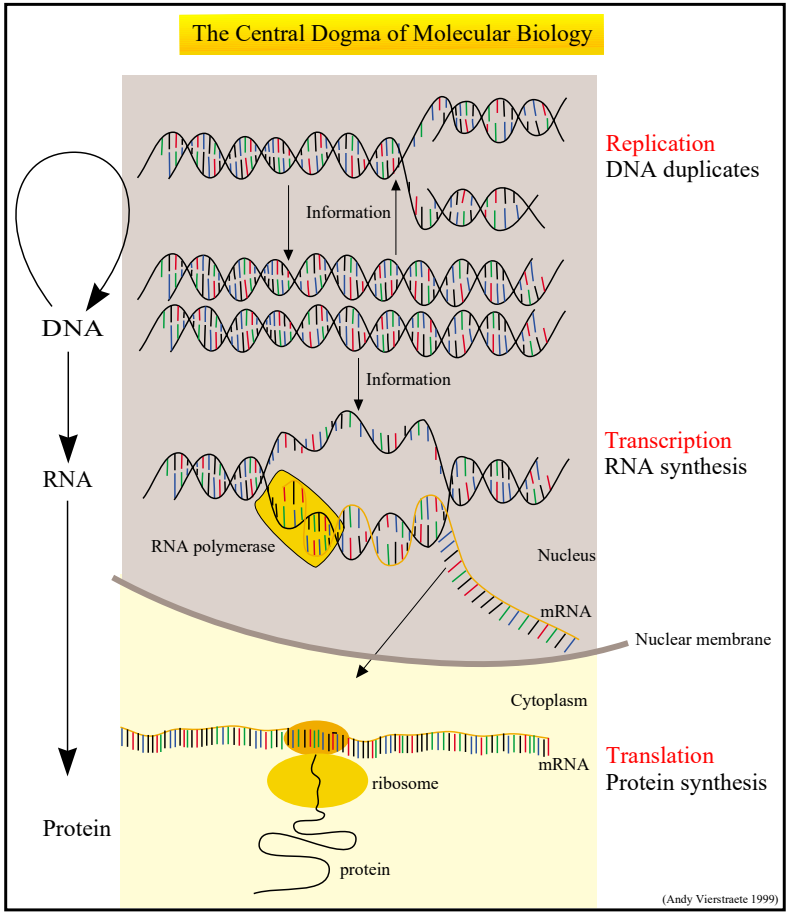


PDB Statistics: Overall Growth of Released Structures Per Year

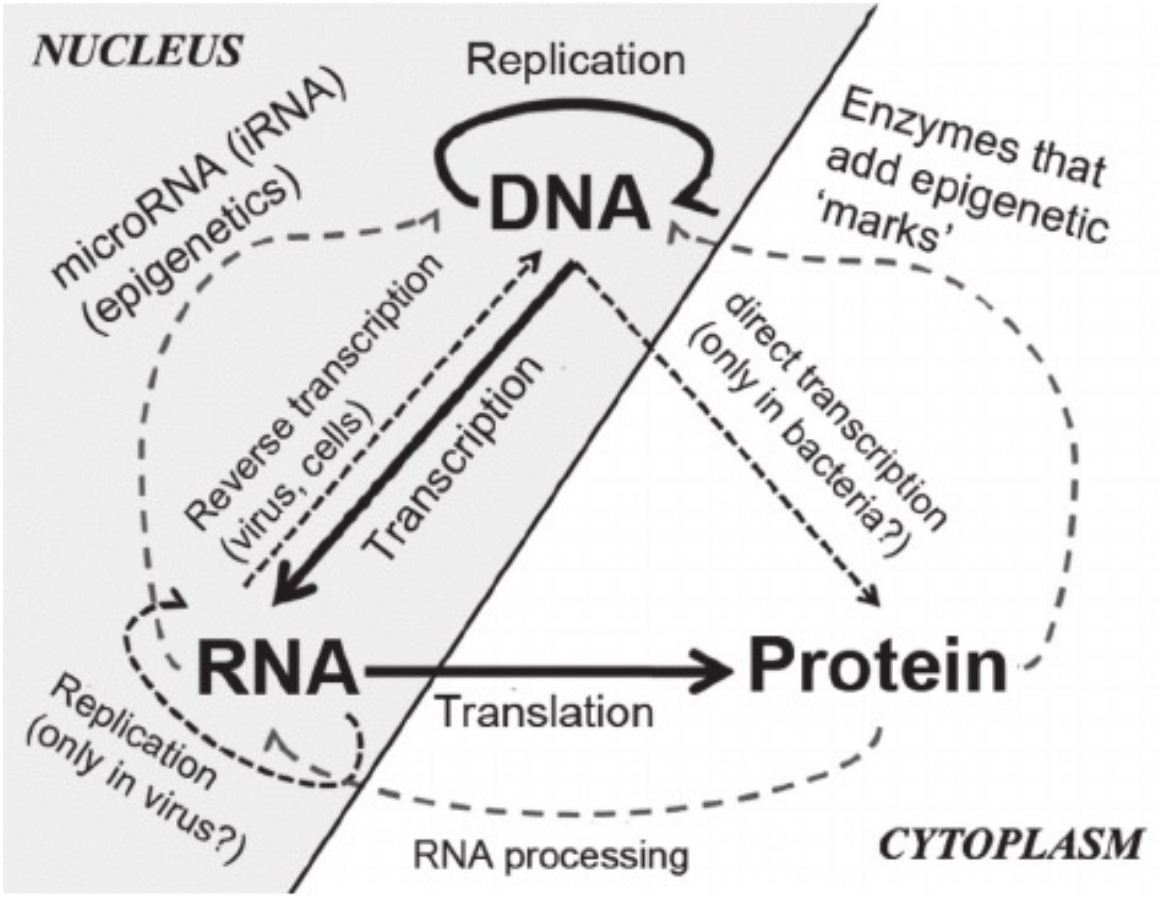


What are ‘omics?



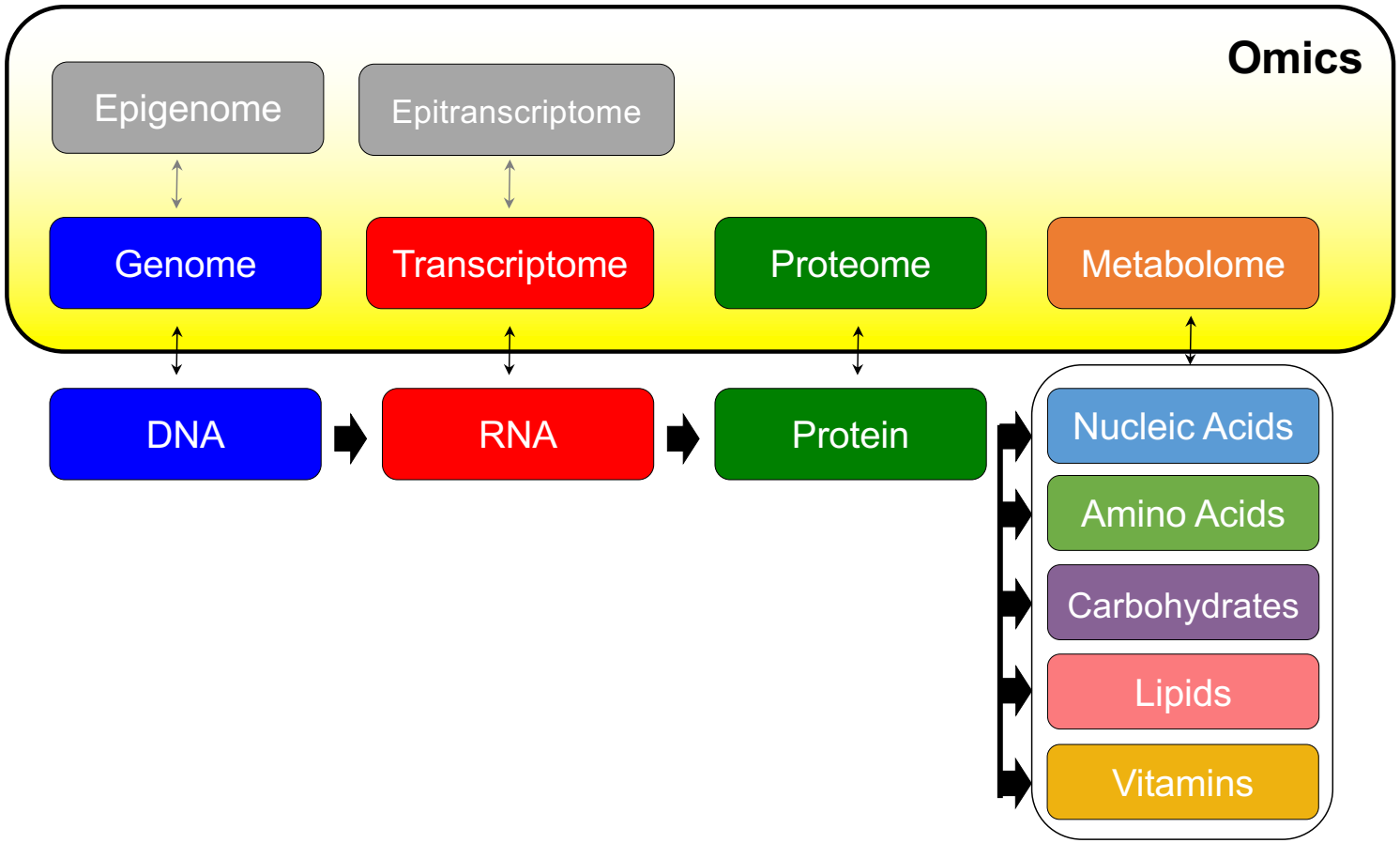


Beyond the Central Dogma



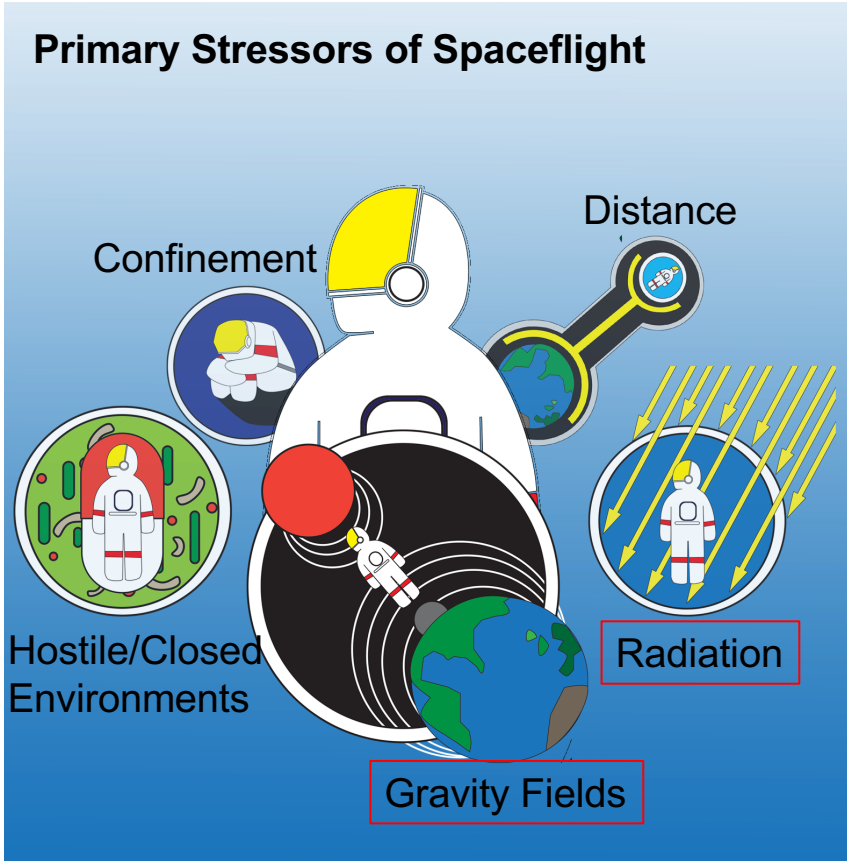
Gonzalez-Pardo and Alvarez, 2013

What are ‘omics?



Why is studying omics important for spaceflight?

- What, when, and where genes are expressed allow for cell type diversity and enable living organisms to respond and adapt to surroundings
- Gene expression is primarily regulated by environmental factors both micro (cell's micro-environment) and macro (organism's external stimuli or stressors)
- Spaceflight alters the transcriptional patterns and molecular signaling networks within our cells, which in turn causes physiological changes
- Understanding such changes will enable development of mitigation strategies to better withstand the rigors of long-duration spaceflight



GeneLab: Open Science for Life in Space (https://genelab.nasa.gov)

Open Science for Life in Space

Home | About | Data & Tools | Research & Resources | Working Groups | Help

Keywords

Welcome to NASA GeneLab - the first comprehensive space-related omics database; users can upload, download, share, store, and analyze spaceflight and spaceflight-relevant data from experiments using model organisms.

Data Repository
Search and upload spaceflight datasets

Analyze Data
Perform large-scale analysis of biological omics data

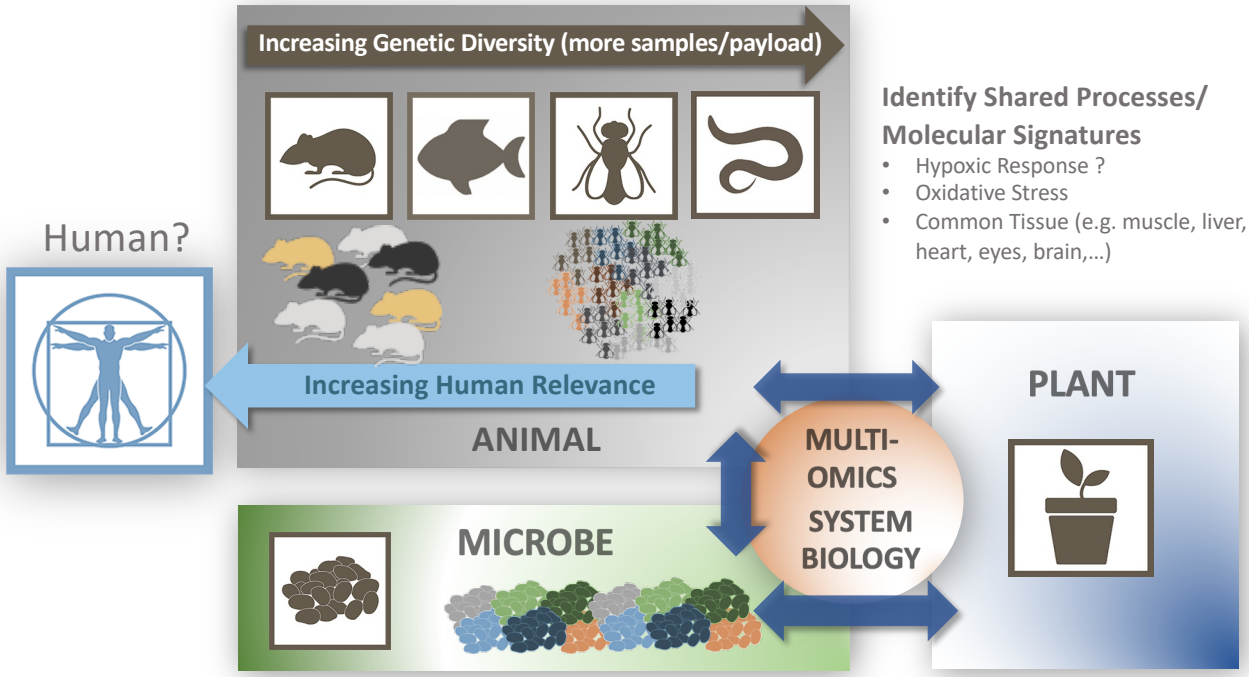
Environmental Data
Radiation data collected during experiments conducted in space

Collaborative Workspace
Share, organize and store files

Submit Data
Have space-relevant data to submit to GeneLab?

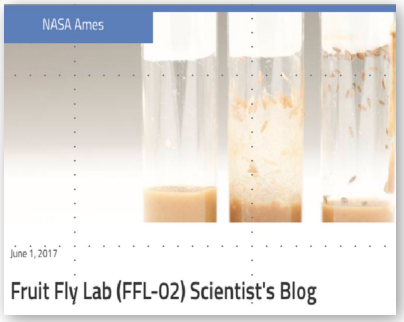
Visualize Data
Interact with GeneLab processed data

GeneLab ecosystem: Maximizing knowledge by bringing experiments together



Identify Shared Processes/ Molecular Signatures

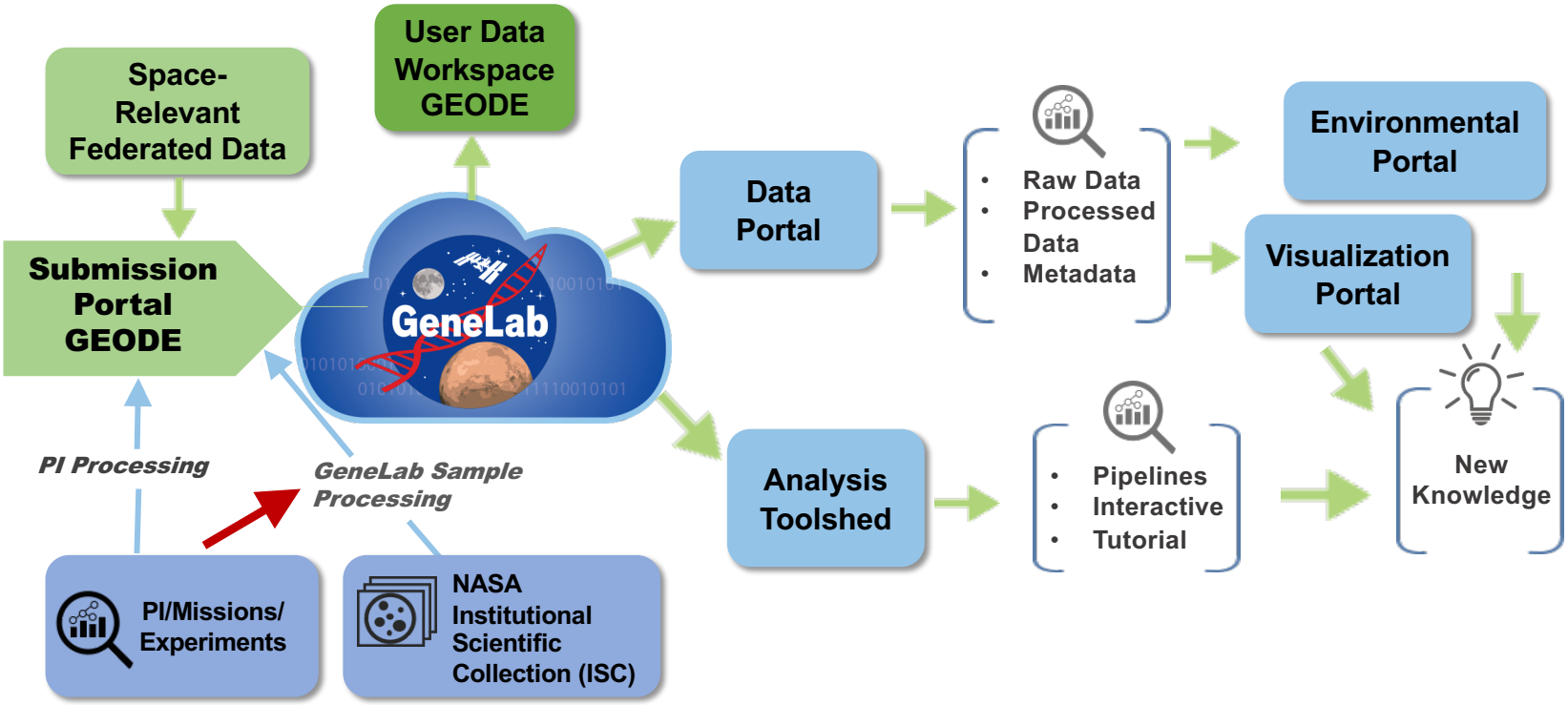
- Hypoxic Response ?
- Oxidative Stress
- Common Tissue (e.g. muscle, liver, heart, eyes, brain,…)



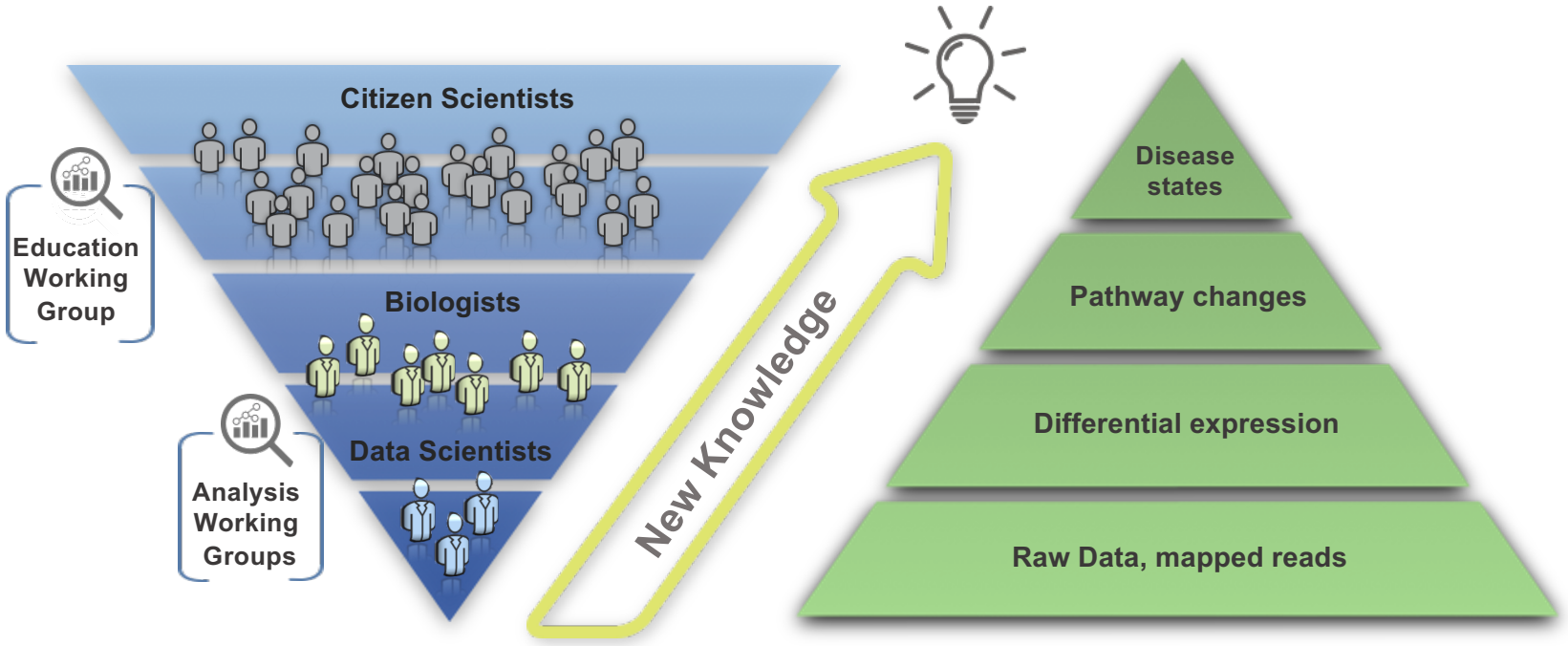
For Spaceflight

- High "n" number – statistically significant data
- Genetically identical animals
- Low resource requirements
- Short life cycle - multiple generations
- Measure response of a whole multicellular animal
- **Flies used as a model for humans for innate immunity, circadian rhythm, oxidative stress, neurobehavior, development, genetics, GWAS, "omics" studies etc**

GeneLab Data System



GeneLab Data Democratization



Data Processing Goals

- Reproducibility
  - Cornerstone of scientific validity
  - Obstacles:
    - Portability across computers
    - Replicating complex network of steps
    - Computational cost

Project Overview

The *Reproducibility Project: Cancer Biology* was an 8-year effort to replicate experiments from high-impact cancer biology papers published between 2010 and 2012. The project was a collaboration between the Center of Open Science and Science Exchange with all papers published as part of this project available in a collection at *eLife* and all replication data, code, and digital materials for the project available in a collection on OSF.

When preparing replications of **193 experiments** from **53 papers** there were a number of challenges.

2% experiments with open data

70% of experiments required asking for key reagents

69% of experiments needing a key reagent original authors were willing to share

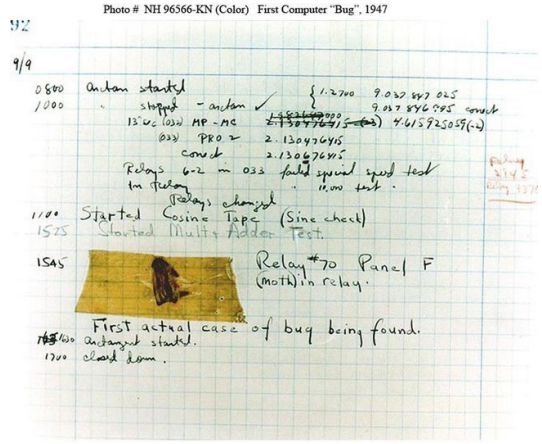
0% of protocols completely described

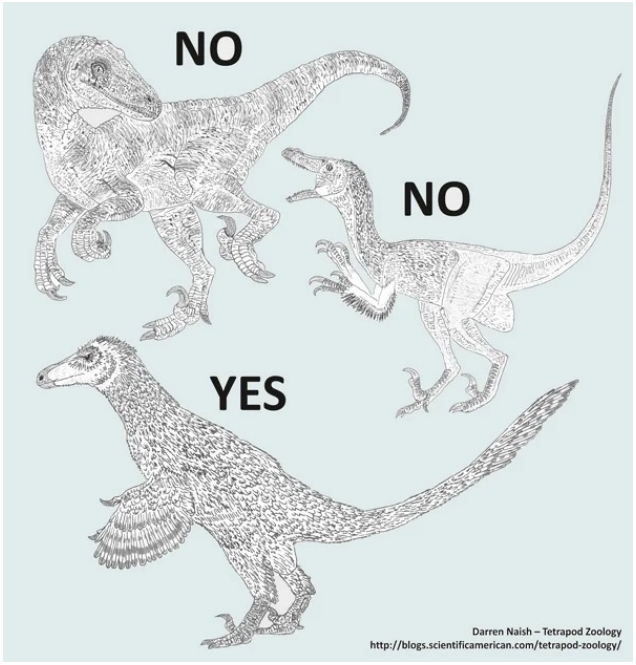
32% of experiments the original authors were not helpful (or unresponsive)

41% of experiments the original authors were very helpful

Data Processing Goals

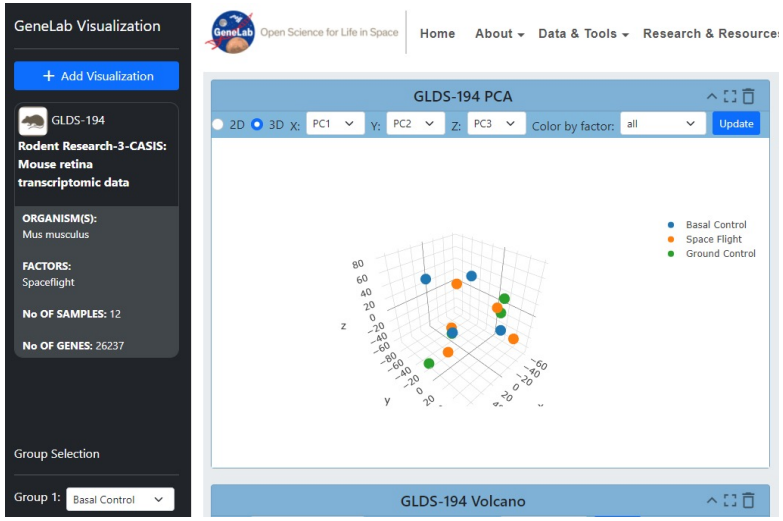
- Data Provenance
  - Maintain a record of data products
  - Ensure confidence in new ideas generated from the data
  - Address newly identify sources of error
    - Science theory and conclusions evolve
    - Software has bugs

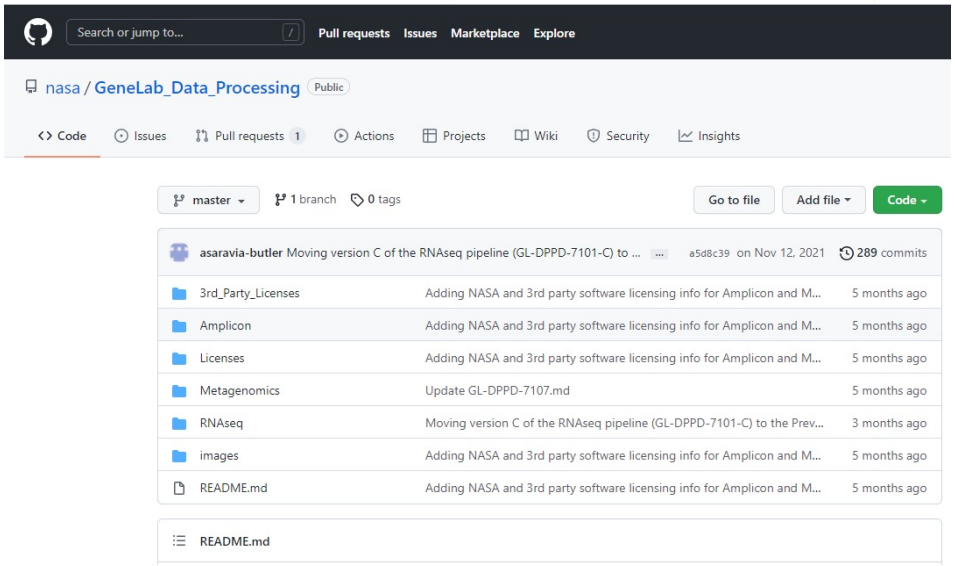




Data Processing Goals

- Extendable Analysis
  - Modularity of steps
  - Entry points along wide range of skill and resource levels





Data Processing Goals: Current Work

Tools I Use

nextflow

GitHub

slurm workload manager

ANACONDA®

Omics Areas of Focus

Current

- Transcriptomics
  - NGS and Microarray

Future:

- Proteomics, Spatial, Single Cell

What I Build

- "One-click" reproducible Nextflow workflows
- Automatic quality flagging software
- Future: New data product extensions

Acknowledgements

- NASA GeneLab Team

- GeneLab Analysis Working Groups (AWGs)
  - Interested in joining an AWG? Visit: https://genelab.nasa.gov/awg/join

- NASA Space Biology Program